Applied Acoustics 159 (2020) 107041

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Environmental sound monitoring using machine learning on mobile devices

Marc Green *, Damian Murphy

AudioLab, Department of Electronic Engineering, Genesis 6, University of York, YO10 5DQ, United Kingdom

ARTICLE INFO

Article history:
Received 18 March 2019
Received in revised form 5 September 2019
Accepted 15 September 2019
Available online 23 October 2019

ABSTRACT

This paper reports on a study to assess the feasibility of creating an intuitive environmental sound monitoring system that can be used on-location and return meaningful measurements beyond the standard L_{Aeq}. An iOS app was created using Machine Learning (ML) and Augmented Reality (AR) in conjunction with the Sennheiser AMBEO Smart Headset in order to test this. The app returns readings indicating the human, natural and mechanical sound content of the local acoustic scene, and implements four virtual sound objects which the user can place in the scene to observe their effect on the readings. Testing at various types of urban locations indicates that the app returns meaningful ratings for natural and mechanical sound, though the pattern of variation in the ratings for human sound is less clear. Adding the virtual objects largely has no significant effect aside from the car object, which significantly increases mechanical ratings. Results indicate that using ML to provide meaningful on-location sound monitoring is feasible, though the performance of the app developed could be improved given additional calibration.

1. Introduction

1.1. Environmental sound and the soundscape approach

In the field of environmental sound monitoring, the prevailing measurement is the L_{Aeq}, which indicates the average A-weighted Sound Pressure Level (SPL) dose received at a measurement location over a period of time [1]. This is simple to understand but does not give any real detail on the content of the sound scene, which can be key in its impact on those experiencing it. Measuring L_{Aeq} creates a flattening effect, in that all sounds are considered to have the same value (or lack thereof). This has been termed the "noise approach" [2], where sound is managed by suppression – reducing absolute levels regardless of source – and is the model followed by the vast majority of legislation covering the issue [3].

More recently, however, drawing inspiration from Murray Schafer's seminal work *The Soundscape* [4], an alternative known as the "soundscape approach" has been emerging. The term 'soundscape' is defined in the ISO12913 standard as the "acoustic environment as perceived or experienced and/or understood by a person or people, in context" [5]. The key to this approach is the idea that human reaction to environmental sound is not uniform, and that the content of the sound sources have a significant effect on this. This leads to the conclusion that some environmental sound should be "perceived as a resource" rather than "managed as a waste" [2]. The soundscape approach therefore requires sound sources to be differentiated in order to be effective, and this is perhaps why it has not seen more widespread adoption. Whilst L_{Aeq} is simple to measure using off-the-shelf devices, gathering data for the soundscape approach has typically involved in situ soundwalks [1,6] or extensive listening tests [7,8], both of which are time-consuming, expensive, and difficult to reliably replicate.

1.2. Sound monitoring using smartphones

Modern mobile smartphone devices have provided a new avenue for environmental sound monitoring, and many apps have been created for this purpose. Most of these apps reflect the noise approach, featuring an implementation of a sound level meter [9,10], sometimes coupled with other environmental measurements such as air quality [11]. A comprehensive list is available in [12]. Some apps created for research projects have used the potential of mobile devices for crowd-sourcing data to create noise maps showing geographical distributions of L_{Aeq} measurements [13–16], similar to the type specified in the Environmental Noise Directive (END) [17]. Maps created using mobile crowd-sensing can potentially be more up-to-date and higher resolution (depending on user engagement) than the simulations typically used to create maps for compliance with the END. It has also been proposed to use noise maps created using smartphone data to suggest noise abatement interventions [18].

* Corresponding author.
E-mail address: marc.c.green@york.ac.uk (M. Green).

https://doi.org/10.1016/j.apacoust.2019.107041
0003-682X/

By contrast, there have been very few apps which use the soundscape approach. The Hush City app [12] seeks to "integrate the soundscape approach with the noise-based one" by creating 'quietness' maps based on sound level measurements in conjunction with a questionnaire that users fill in at test locations. This is certainly a step towards incorporation of the soundscape approach, but the use of questionnaires is subject to some of the same problems previously outlined in relation to listening tests and soundwalks.

1.3. Machine learning for sound monitoring

There has recently been research into using Machine Learning (ML) techniques to analyse and identify sounds in everyday acoustic environments. Whilst most previous work using ML in audio focuses on speech recognition or music analysis, the recent series of DCASE (Detection and Classification of Acoustic Scenes and Events) challenges [19] have been the focal points of a large increase in research for using ML to identify everyday sounds. The EigenScape database [20] was created specifically to provide a basis for development of ML techniques for soundscape analysis.

While use of ML on smartphones for speech recognition and music identification is widespread, there have been very few apps designed to conduct environmental sound recognition. Apps by Cordeiro and Barbosa [21], and Lu et al. [22] classify incoming sounds as either speech, music or 'environmental'. These classes have limited use for the soundscape approach, however, as all environmental sound is conflated in a manner essentially similar to the noise approach. Lane et al. [23] created a classifier to run on mobile devices that categorised environmental sound as either music, traffic, voicing or 'other'. This could be more useful for the soundscape approach, but the classifier has not been implemented in any available app.

1.4. Augmented reality audio

On the cutting edge of current smartphone technology is Augmented Reality (AR), whereby virtual objects are superimposed onto a live camera feed of the real world environment. Apple's ARKit [24] can track features in the device's surroundings to enable a smooth AR experience, and is emerging as a viable tool not only for gaming, but also for interior design and measurement applications such as IKEA Place [25] and Housecraft [26].

The Sennheiser AMBEO Smart Headset (ASH) [27], shown in Fig. 1, is an accessory for iOS devices that can be used to extend AR to the audio domain. The ASH features microphones built into each earpiece, which can be used to record binaural audio. The 'Transparent Hearing' mode allows incoming audio to be relayed instantly to the in-ear speakers. This can be blended with audio from the device to create an augmented audio scene in a similar manner to ARKit's handling of visuals.

Whilst spatial audio, often used in conjunction with Virtual Reality (VR), is an established delivery format for auralisations of soundscapes [28,8], few studies have been done which incorporate AR audio. This is despite the suggestion from Hong et al. that it could be "useful for projects that involve altering the soundscapes of existing locations...[enabling] soundscape researchers to fuse the virtual sound sources seamlessly with real sound" [28]. Kınayoğlu [29] created a system to test the perceptions of subjects to altered on-location acoustic scenes, replacing local sound with spatial soundscapes created using recordings from other locations. Whilst this system featured head-tracking for realistic sound spatialisation, this was not true AR as the existing location sound was completely overlaid by the virtual sound scene – there was no microphone component in this system to create a blend of real and virtual audio.

Fig. 1. The Sennheiser AMBEO Smart Headset, showing earpieces with binaural microphones, control unit and connector.

1.5. Aims and objectives

This study seeks to find whether it is feasible to create an intuitive measurement system for environmental sound monitoring that runs on a handheld device and uses machine learning techniques to provide meaningful readings beyond the standard L_{Aeq}. These readings should have relevance to human soundscape perception. To this end, an iOS app was created that uses the ASH combined with machine learning technologies to provide a more nuanced measurement of environmental audio in accordance with the soundscape approach.

An AR component of the app was also designed in order to test the usefulness of the output from the ML component in terms of assessing interventions that might be added to the environment to affect its soundscape. The app allows users to place virtual objects, having both a sonic and visual component, into the environment. These can be moved and altered by the user, with the augmented scene available for listening and also passed to the ML component for analysis.

The App was developed with two goals in mind:

1. Provide a simple interface for the measurement of acoustic environment properties beyond L_{Aeq}.
2. Using AR technology, allow users to test the effects on these measurements of potential alterations to the environment.

There are clear applications for this kind of app in soundscape research, but also more broadly in urban planning, where AR could assist with exterior design, or testing proposed alterations of public spaces.

2. App development

2.1. Soundscape taxonomy

Recent research into soundscape perception [7,8] has used three main groups of environmental sound sources:

- **Natural**: The sounds of all manner of fauna except humans, together with sound created by weather and geological forces including rainfall, wind and flowing water.

- **Mechanical**: Sounds from machinery, including transport and construction.
- **Human**: Non-mechanical sounds indicative of the presence of humans. This primarily consists of speech, but also footsteps, music and laughter.

Some previous work (with origins in soundecology and biodiversity research) [30–32] uses an alternative taxonomy, classifying sounds as *anthrophony*, which broadly speaking groups human and mechanical sounds together, or *biophony* and *geophony*, which split natural sounds into those produced by animals and those produced by geological forces. Whilst this taxonomy is no doubt useful for soundecology applications, a great deal of research on human soundscape perception has shown most responses are dependent on two components, sometimes labelled *pleasantness*, most affected by the natural/mechanical balance, and *eventfulness*, mainly dependent on the presence of human sounds [33–35]. This is reflected by the use of *Valence* (positive/negative emotional state) and *Arousal* (apathetic/excited emotional state) assessment scales in [8]. It was therefore decided that the app should display ratings for natural, mechanical, and human sounds, which could be used to estimate *pleasantness* and *eventfulness*.

2.2. Core ML model creation

Since we are interested in the overall content and character of sound scenes in general, rather than on detection of individual sources in particular, we used an Acoustic Scene Classification (ASC) framework [36] for the ML models. The usual goal of ASC is for the model to assign a label to incoming audio clips indicating the class of location the clip was recorded in. In this work, specific scene classifiers were reappropriated to provide estimates for the prevalence of the human, natural, and mechanical components of scenes.

Apple's Core ML library [37] was used to create an object within the app that performs analysis on the audio incoming from the ASH. Core ML includes a tool that can translate certain models created using the *scikit-learn* Python library [38] into an iOS-compatible format.

Models were trained using audio from the EigenScape database [20]. Mel-Frequency Cepstral Coefficient (MFCC) features were extracted from the zeroth-order (mono-omni) channel in a manner similar to the baseline models in [19,20]. Classifiers were trained for all eight location classes present in the EigenScape database (Beach, BusyStreet, Park, PedestrianZone, QuietStreet, ShoppingCentre, TrainStation and Woodland). Since EigenScape features eight examples of each location class, models were trained on six recordings and tested on the remaining two. In [20], MFCCs were extracted using the *librosa* library [39], but since this app requires MFCC features to be extracted on the iOS device in real-time, the *aubio* library [40] was used as an alternative, as it is compatible with both iOS and Python. The library was configured to extract 20 MFCC coefficients, covering the frequency range up to approximately 11 kHz. In [20], Gaussian Mixture Models are used to classify sound, whilst this work uses Support Vector Classifiers (SVCs) for compatibility with Core ML. Features were extracted from frames of 2048 samples using rectangular windows with no overlap, resulting in 84,375 training frames for each class.

Fig. 2 shows the performance of the eight models in a confusion matrix. It can been seen that whilst the models for BusyStreet and Woodland perform well, models for the other scenes were generally inaccurate. This was not so much a problem for this work, however, as the primary interest here is in reporting of alternative metrics for sound scenes, rather than precise scene classifications.

From these results, the BusyStreet classifier was chosen to provide mechanical ratings, with the Woodland classifier chosen for natural ratings. The prevalence of vehicle sound in the BusyStreet scenes and birdsong in most Woodland scenes make them largely representative of these sound categories, an assumption reinforced by listening tests conducted in [7]. Choosing a model for human ratings was less simple, as the most obvious classifier – PedestrianZone – did not perform accurately. The ShoppingCentre classifier was instead chosen for this purpose as, whilst only successful at identifying 50% of the ShoppingCentre scenes, misclassifying TrainStation scenes the remaining 50% of its output, both of these scenes have a relatively large human sound component.

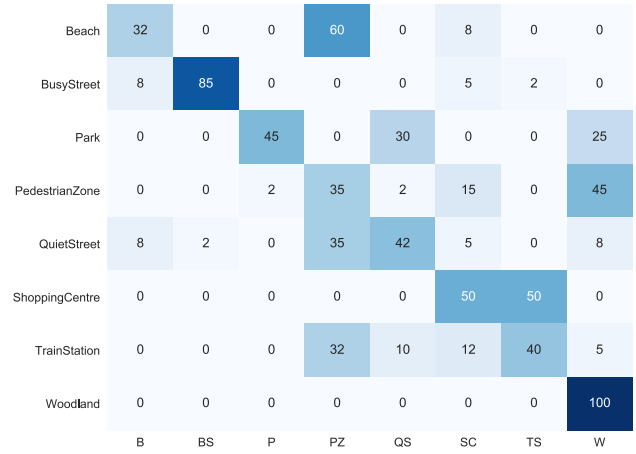


Fig. 2. Confusion matrix showing performance of the aubio-extracted MFCC-trained SVCs (percentage correct classifications).

Each of these models produces a rating indicating the probability that MFCC features extracted from incoming audio frames came from an acoustic scene similar to those they were trained on. In [20], the model returning the highest probability is used to generate a scene label. In this app, these probabilities are reappropriated as ratings for each sound source group, which are displayed to the user. In essence, we obtain estimates for the three components by measuring the similarity of the incoming audio to the three chosen scene models.

2.3. AR audio implementation

2.3.1. AR audio sources

In order to implement AR audio as well as visuals, custom objects were required to couple 3D graphics with realistic audio sources using binaural processing. Apple's SceneKit objects have a built in audio player instance for "3D audio" [41], but in testing it was found these use standard stereo panning only. Apple's audio framework (AVFoundation) does, however, include an object called the AVAudioEnvironment node, which features an option to use high-quality Head-Related Transfer Function (HRTF) rendering for binaural output. Our custom object therefore adds an audio player to the standard SceneKit node object, with the 'position' parameter of the audio set to mirror the visual position of the node.

2.3.2. AR acoustic barrier object

In addition to AR audio sources, an object was created to simulate the addition of an acoustic barrier to a scene. Acoustic barriers are a fairly common noise abatement intervention in deployment along the side of roads or railway lines [42]. Sound is attenuated primarily by diffraction - the barrier blocks the direct path, so sound must travel over the top to reach the receiver. The path length difference δ is critical to the attenuation performance of the barrier, and is calculated as the difference between the length of the diffracted path from source to receiver (over the barrier) and

the blocked direct path. Eq. (1) shows how attenuation A varies with δ and sound wavelength λ [42].

$$A\,[\mathrm{dB}] = 10\log_{10}\left(3 + \frac{40\delta}{\lambda}\right) \tag{1}$$

The result of this is that the larger the path length difference, the greater the attenuation, with high frequencies attenuated more than low frequencies [42,43]. In practise, this means that barriers are most effective when placed close to the sound source or receiver.

To simulate the effect of adding a sound barrier to the scene, our virtual barrier selectively filters the real-world sound picked up by the ASH before this is relayed to the listener as part of the complete augmented audio mix. This is achieved by using a stereo low-pass filter (LPF), blending with the dry signal from the ASH mics and panning its output with respect to the angle between the listener and the barrier.

With regards to calculating the path length difference, there is no way at present to measure the distance between the virtual barrier object and the various sound sources making up a real-world scene, however the distance to the receiver (listener) is known. The cutoff of the LPF, representing the amount of high-frequency attenuation provided by the barrier, is therefore calculated based on the distance between the camera position and the virtual barrier. The cutoff is set at 20 Hz if the user is directly next to the barrier, and reaches 20 kHz once the user moves 10 metres away, effectively neutralising the filter's perceptual effect and mimicking the negligible impact of real-world sound barriers given very small path length differences. This gives a reasonable illusion of the attenuation of high-frequency sound incoming from a certain direction as the user turns the camera and moves around in the scene. Future versions of this app could incorporate more sophisticated models of barriers and outdoor sound propagation as defined in ISO 9613 [44].

2.4. App structure

2.4.1. User interface

The various interface elements of SoundscapeAR are shown in Fig. 3. The main window of SoundscapAR (Fig. 3a) shows the live camera feed and any active virtual objects. There are three subviews performing various functions that can be shown and hidden by the user using the three small buttons in the lower right of the interface.

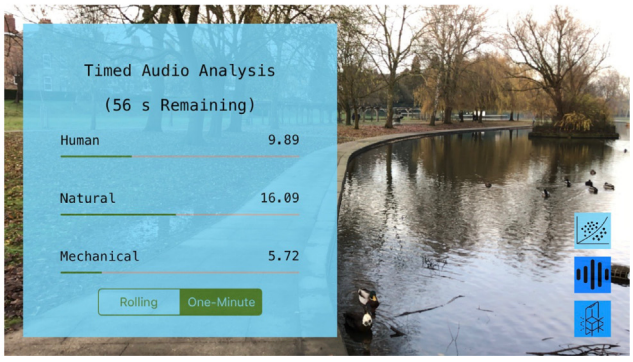
The AR status window is visible on startup and indicates whether ARKit has detected a plane. Detection of a real-world horizontal flat surface (usually corresponding to the floor) is necessary before ARKit is able to properly track the environment. Once the plane is detected, a text indicator turns green and the window becomes redundant. The user can now proceed to place objects.

The AR objects window is shown in Fig. 3b. There are four virtual objects available for the user to place – car, bird, water fountain and barrier. These are represented by four icons that show red or green to indicate whether each object is active. This view also shows crosshairs over the live camera feed. Tapping on each icon places the corresponding object into the virtual scene at the position on the detected plane indicated by the crosshairs. The user can then drag the virtual object to fine-tune the positioning, if desired. Tapping on the icon again removes the object from the scene.

The Audio analysis window, pictured in Fig. 3c, shows the output probabilities from the Core ML object. By default, the window shows a 1-s rolling average. The user can also select a 1-min average recording mode similar to that used by SPL meters to record L_{Aeq}.

(a) Main app view.

(b) Object selection window.



(c) Audio analysis window.

Fig. 3. Various views of the SoundscapAR interface, showing virtual objects added to real scenes, as well as the user interface for adding/removing objects, and the readout of soundscape parameter ratings.

2.4.2. Audio flow

Fig. 4 shows the structure of the audio signal flow through the SoundscapAR app. The binaural audio input from the ASH is filtered by the stereo LPF if the barrier object is active before being mixed with audio from any active virtual sources. The main mixer output is then passed to the ASH speakers.

An MFCC feature extractor powered by *aubio* processes frames of 2048 samples sourced from a tap applied to the main mixer output. The extracted MFCCs are sent to the Core ML object, which returns probability ratings for human, natural, and mechanical audio sources in near real-time. This process is illustrated by the dotted lines in Fig. 4. In this way, with all virtual objects disabled, the user can record 'clean' ratings for an acoustic scene. Virtual objects can then be placed and the scene re-analysed to observe any effect on the ratings the added objects may have.

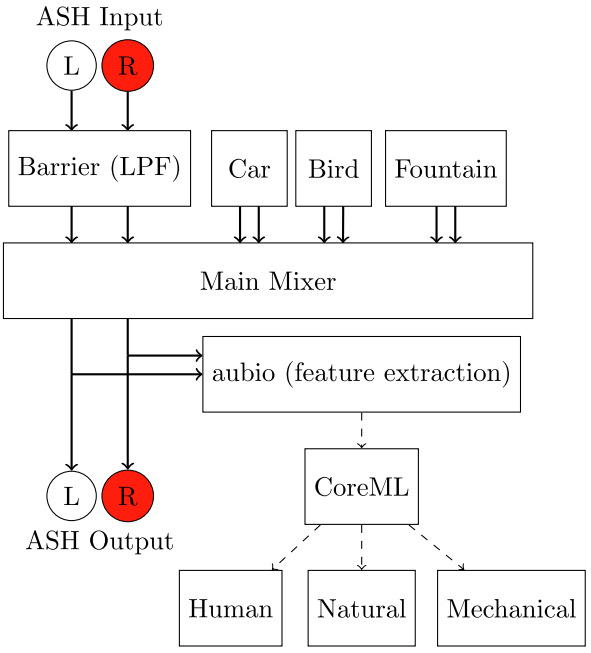


Fig. 4. Diagram showing the complete audio flow in SoundscapAR. Dotted lines indicate numeric (non-audio) values.

3. Testing

3.1. Methodology

To test the effectiveness of the app for environmental sound monitoring and the effects of the virtual objects, the app was loaded to an iPhone 7 and taken to 6 locations around the city of York in the UK. These locations, mapped in Fig. 5, were chosen to represent a good variety of urban environments, including busy streets (Bishopthorpe Road, Exhibition Square), pedestrian areas (Shambles Market), more natural areas (Rowntree Park), and locations that combine these characteristics (York Piccadilly, Tower Gardens).

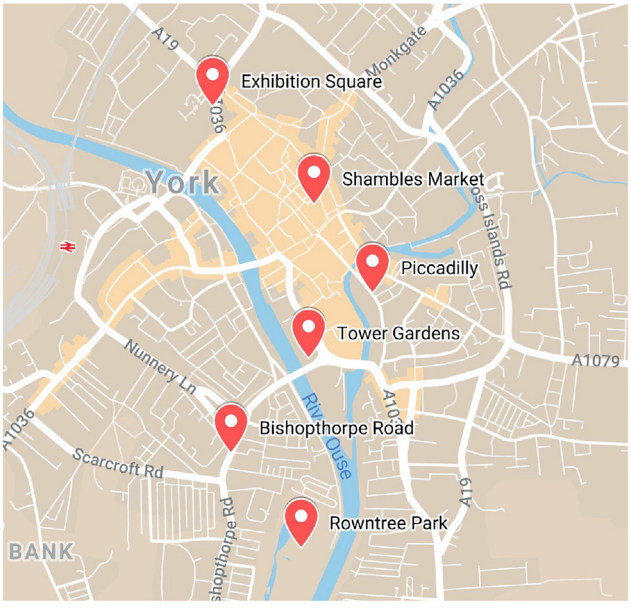


Fig. 5. Map showing the locations at which the app was tested.

The audio analysis feature was used to record repeated one-minute average ratings at each location with various virtual objects added as follows:

- No virtual objects (clean reading)
- Barrier
- Bird
- Car
- Fountain
- Barrier/Bird/Fountain

Objects were placed a reasonably realistic distance in front of the listener location - generally between 2 and 4 meters. In the multi-object condition, the barrier and the fountain were placed on opposite sides of the listener location, with the bird placed roughly above the listener. Using these readings, the classifier's effectiveness in terms of delivering plausible and useful ratings for each location can be investigated. The effect of adding each virtual object can also be tested, as well as whether adding multiple objects has any cumulative effect.

3.2. NDSI/pleasantness rating

One of the key advantages of the L_{Aeq} is its simplicity in interpretation, in that it distils complex sound scenes into a single number, albeit one that is not useful for the soundscape approach. The field of soundecology has proposed several alternative metrics that might be more useful for the soundscape approach, yet still be simple to understand. One of these is the Normalised Difference Soundscape Index (NDSI), which is intended to "estimate the level of anthropogenic disturbance on the soundscape by computing the ratio of human-generated to biological acoustic components" using a scale of ± 1 [32,45]. In our formulation, for increased perceptual relevance we substitute anthrophony for mechanical sounds (-1) and biophony for natural sounds (+1). Our version therefore could be thought of as a metric describing the *pleasantness* dimension of soundscape perception.

In [32,45], the NDSI value is estimated by finding the ratio between the power spectral density of the 1 kHz - 2 kHz band (said to be more prevalent in mechanical sound) and the 2 kHz - 11 kHz band (said to be more prevalent in natural sound). This rudimentary approach results in unreliable output, though the shortcoming is noted in [45], which states "advancements are needed to help characterise and search acoustic observations". A machine learning model such as the one employed in this app could represent just such an advancement.

To test the response of the system and its viability as a robust way to calculate an NDSI/pleasantness metric, natural and mechanical ratings from each location (with and without virtual objects present) were used to calculate ratings as follows [32]:

$$\mathrm{NDSI} = (\beta - \alpha)/(\beta + \alpha) \tag{2}$$

where α and β are the reported mechanical and natural ratings, respectively.

4. Results

Fig. 6 shows the NDSI/pleasantness values for each scene. The outliers shown are the measurements recorded with the virtual car present (see results in Section 4.1). Rowntree Park has the highest value, followed by Shambles Market, and then Tower Gardens. This shows the effectiveness of the classifier as both the park and the market are low in mechanical sounds, though there is some quieter machinery present at the market (small generators etc.). Tower Gardens is nearer to a main road, and the lower value

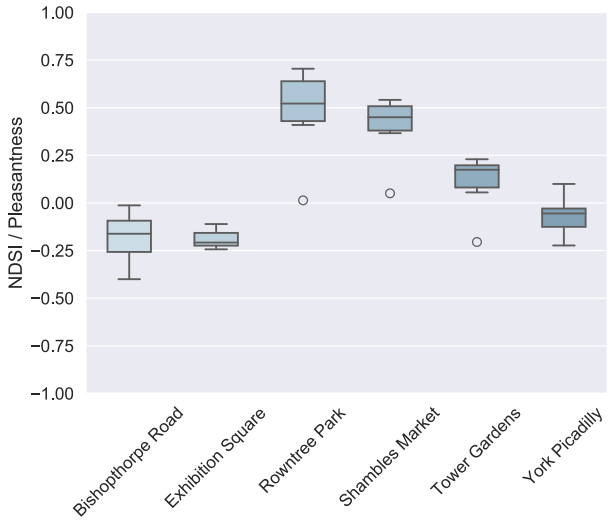


Fig. 6. NDSI/pleasantness values for each recording location. Positive ratings indicate prevalence of natural sound, whereas negative ratings indicate prevalence of mechanical sound.

reflects this. Bishopthorpe Road and Exhibition Square both have heavy traffic, and this is reflected in that their values are the lowest. Piccadilly has slightly lighter traffic, and values are slightly higher in general.

These results do reveal a skew towards the upper end of the scale. Bishopthorpe Road and Exhibition Square, which have heavy traffic, value further towards the middle of the range than might be expected. The mean mechanical value overall is $8.93 \pm 1.32\%$, whereas the mean natural rating is $12.87 \pm 1.94\%$. Given the breadth of locations chosen, these should ideally be more similar. This suggests the models require some calibration.

The trio of ratings gathered for each scene with no virtual objects present is shown in Fig. 7. It can be seen here than the human rating does seem to give some additional information beyond the two poles of the NDSI/pleasantness metric. Exhibition Square, for instance (7b) has a similar human rating to Shambles Market (7d), whereas their other ratings vary greatly. Despite this, the human ratings clearly do not vary as much from place to place as the others – the variance in human ratings is 6.99, where variance in mechanical ratings is 15.35 and natural variance is even higher, at 24.79. It is unclear whether this is a flaw in the classifier, or whether variation in human sound is smaller than the other categories in the locations investigated.

4.1. Effect of virtual objects

Fig. 8 shows the distributions of human, natural, and mechanical ratings from each scene plotted against the activation of various virtual objects. The data was analysed using D'Agostino's K^2 test [46], which indicated normal distributions for all three sets of ratings. It can be seen in Fig. 8a that human ratings hover around a mean of 15 % regardless of objects added. Repeated measures ANOVA shows no significant effect of adding any object $F(4, 20) = 1.49, p = 0.24$.

Natural ratings (Fig. 8b) show more of a spread than human ratings in general and seem impacted somewhat by the addition of the car object. This reduced the mean rating from 12.83 ± 4.98 % to 10.20 ± 2.06 %, whilst adding the fountain increased the mean to $14.96 \pm 5.14\%$. Repeated measures ANOVA here shows the effect of adding objects on the natural rating is significant, $F(4, 20) = 5.06, p < 0.05$. Post-hoc paired t-tests using the bonferroni correction show no individually significant contributors.

(a) Bishopthorpe Road (b) Exhibition Square

(c) Rowntree Park (d) Shambles Market

(e) Tower Gardens (f) York Piccadilly

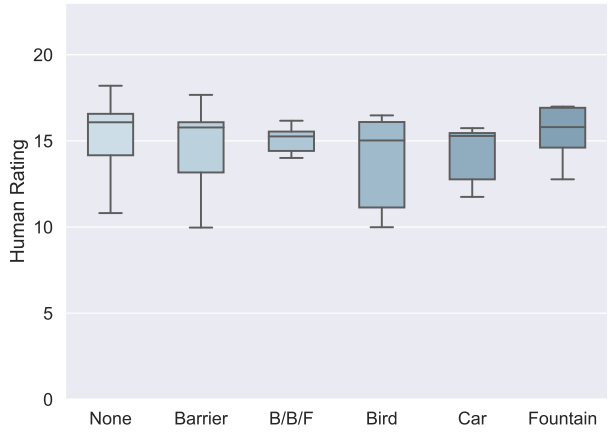
Fig. 7. Radar plots showing the three ratings gathered for each location with no virtual objects added.

The biggest effect recorded was on the mechanical ratings by the addition of the car object, as can clearly be seen in Fig. 8c. The mean rating increases from $8.93 \pm 3.92\%$ to $13.69 \pm 1.79\%$. Repeated measures ANOVA shows significance, $F(4, 20) = 10.79, p < 0.05$. Bonferroni-corrected post hoc testing shows a significant effect on the ratings from the car object ($t(5) = 4.24, p < 0.0125$), but no significant effect from any other objects.

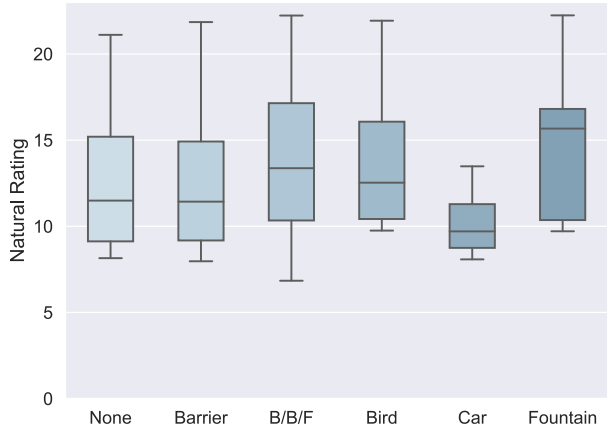
4.2. Discussion

Generating the NDSI/pleasantness metric using natural and mechanical ratings produced some plausible results, with values that matched location characteristics well. This suggests that a machine learning approach to calculating meaningful soundscape indices could be effective, and that such a system could be incorporated into an easy-to-use handheld device. In future work it would be interesting to compare the NDSI/pleasantness values obtained here to results from the original frequency-ratio method of calculation, and to ratings of these sound scenes by subjects in a listening test. A future version of this app could aim to feature a pleasantness/eventfulness visualisation instead of, or in addition to, the three ratings presented here, though improvements to the human classifier may be required before it can be considered a reliable estimator of eventfulness.

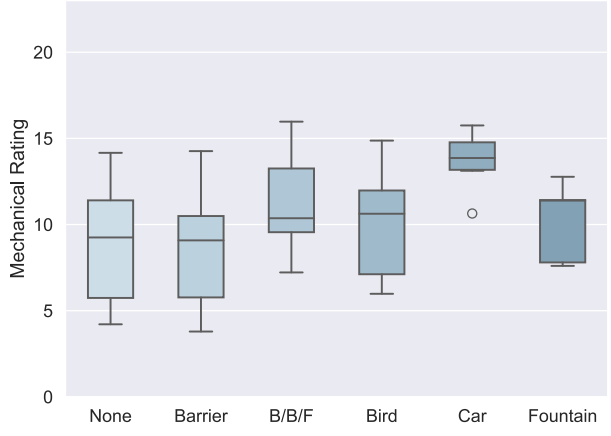
The results from the natural and mechanical classifiers show that these classifiers are to some extent successfully generalising



(a) Human ratings



(b) Natural ratings



(c) Mechanical ratings

Fig. 8. Boxplots showing distributions of human, natural and mechanical ratings across all locations under each virtual object condition.

to audio that is not contained within the EigenScape dataset used for training. In [20] the classifiers are tested on recordings from the same dataset, made using the same equipment. In this study, however, the classifiers are tested at locations not recorded in EigenScape and using the ASH microphones rather than the Eigenmike array [47] which was used to record the EigenScape dataset.

Despite the limited amount of data obtained, there is some indication that the addition of the virtual car tends to cause an increase in the mechanical rating, with a corresponding slight decrease in the natural rating. Addition of 'natural' sources seem to have a very modest effect in increasing natural ratings, and no consistent effect on the mechanical ratings. Addition of the barrier object has very little effect at all on any of the ratings, suggesting the barrier in principle or in the implementation described here (see Section 2.3.2) is not effective. This is possibly due to MFCCs extracted from lower frequencies providing more discriminative information to the classifiers than those from higher frequencies that are more attenuated by the barrier. It is possible that if listening tests were conducted, the barrier object might be rated as perceptually more effective in altering the sound scene than is apparent here.

The fact that the introduction of the virtual car has a much more pronounced effect on the ratings than any of the natural objects aligns with findings presented by Stevens in [7], where the addition of a single car to a sound scene recorded by a lake caused a large increase in mechanical ratings provided by subjects in a listening test. This provides some evidence that the natural and mechanical classifiers produce ratings that are somewhat aligned with human perception, though more study would be needed to corroborate this.

None of the virtual objects seem to have much effect on the human ratings. This is possibly due to the fact that none of the virtual objects implemented could be considered human sound sources. A virtual 'conversation' object might have been more effective in this regard. On the other hand, since the human ratings are less variable generally than the natural and mechanical ratings, it could be that the classifier is not as effective as those trained to identify natural and mechanical sources.

5. Further work

The clear next step with this work would be conducting subjective listening tests with real users interacting with the app's augmented audio. Their ratings could be compared with the classifier outputs in order to reinforce or disprove the results obtained. Indeed, more robust classifiers might be obtained by including listening test results as part of the training stage. This method, explored previously in [34], would perhaps be more robust than re-appropriating a scene classification system, as in the present work.

The classifiers used here could be further improved by utilising more advanced audio features. The MFCC features used here are basic, and it is shown in [20] that spatial audio features can outperform them for scene classification applications. Spatial features could be derived from the ASH's binaural input, but since feature extraction must happen on-device in near real-time, processing power could become a bottleneck in this regard.

The implementation of the app's virtual objects could also be improved. At present, all objects are stationary point sources. Some sources (e.g. the car) would in reality likely be in motion, and some sources would be diffuse. It should be possible to implement these features in a future version of the app. It might also be possible to use more sophisticated processing to make the effects of the virtual barrier more realistic. Like any improvements in feature extraction, however, this would have to take into account the limited processing power available on the device.

Perhaps the most exciting future development could be built upon the "persistent experience" feature introduced in ARKit 2 [48]. This allows AR apps to be "experienced by multiple users simultaneously, and resumed at a later time in the same state". This creates the possibility of conducting AR soundwalks, where virtual objects are placed by a researcher in advance and participants can explore the AR audio environment live. This could be a powerful tool for future research and urban planning.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding was provided by a UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award and the Department of Electronic Engineering at the University of York.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apacoust.2019.107041.

References

[1] Harriet S, Murphy DT. Auralisation of an urban soundscape. Acta Acustica united with Acustica 2015;101(4):798–810. https://doi.org/10.3813/aaa.918874 (Jul 2015).
[2] Brown AL. Soundscapes and environmental noise management. Noise Control Eng J 2010;58(5):493–500. https://doi.org/10.3397/1.3484178 (2010).
[3] Noise – Environment – European Commission, http://ec.europa.eu/environment/noise/index_en.htm [accessed January 8, 2019].
[4] Schafer RM. The soundscape: our sonic environment and the tuning of the world. Inner Traditions/Bear & Co 1993;1993.
[5] International Standards Organisation, ISO 12913–1:2014 - Acoustics – Soundscape – Part 1: Definition and conceptual framework; 2014.
[6] Brooks B, Schulte-Fortkamp B. The soundscape standard. Internoise 2016;2016 (2016):2043–7.
[7] Stevens F, Murphy D, Smith SL. Soundscape categorisation and the self-assessment manikin. In: Proceedings of the 20th International Conference on Digital Audio Effects (2017).
[8] Stevens F, Murphy D, Smith SL. Soundscape auralisation and visualisation: A cross-modal approach to soundscape evaluation. In: Proceedings of the 21st International Conference on Digital Audio Effects (2018).
[9] Aircasting, https://play.google.com/store/apps/details?id=pl.llp.aircasting&hl=en_GB [accessed: 2019-07-08].
[10] The noise app, https://apps.apple.com/gb/app/the-noise-app/id926445612, [accessed: 2019-07-08].
[11] The noise app, https://play.google.com/store/apps/details?id=fr.inria.mimove.quantifiedself&hl=en_GB [accessed: 2019-07-08].
[12] Radicchi A, Henckel D, Memmel M. Citizens as smart, active sensors for a quiet and just city. the case of the "open source soundscapes" approach to identify, assess and plan "everyday quiet areas" in cities. Noise Mapping 2017;4 (1):104–23 (2017).
[13] Zappatore M, Longo A, Bochicchio MA, Zappatore D, Morrone AA, Mitri GD. Mobile crowd sensing-based noise monitoring as a way to improve learning quality on acoustics. In: International Conference on Interactive Mobile Communication Technologies and Learning (IMCL), Thessaloniki, Greece.
[14] Sakagami K, Satoh F, Omoto A. Revisiting acoustics education using mobile devices to learn urban acoustic environments: recent issues on current devices and applications. Urban Sci 2019;3(3):73 (Jul 2019).
[15] Maisonneuve N, Stevens M, Ochab B. Participatory noise pollution monitoring using mobile phones. Inf Polity 2010;15(2010):51–71.
[16] Zuo J, Xia H, Liu S, Qiao Y. Mapping urban environmental noise using smartphones. Sensors 2016;16(10):1692 (Oct 2016).
[17] Directive 2002/49/ec of the european parliament and of the council of 25 june 2002 relating to the assessment and management of environmental noise, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32002L0049&from=EN [accessed: 2019-07-08].
[18] Zappatore M, Longo A, Bochicchio M, Zappatore D, Morrone A, De Mitri G. A mobile crowd-sensing platform for noise monitoring in smart cities. EAI Endorsed Trans Smart Cities 2016;1(1):151627.
[19] Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD. Detection and classification of acoustic scenes and events. IEEE Trans Multimedia 2015;17 (10):1733–46. https://doi.org/10.1109/TMM.2015.2428998 (October 2015).
[20] Green MC, Murphy D. Eigenscape: a database of spatial acoustic scene recordings. Appl Sci 2017;7(11):1204. https://doi.org/10.3390/app7111204 (Nov 2017).
[21] Cordeiro J, Barbosa A. Using smartphones as personal monitoring tools for the acoustic environment. In: Tecnoacustica, Murcia, Spain, 2013.
[22] Lu H, Pan W, Lane ND, Choudhury T, Campbell AT. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In: 7th Annual International Conference on Mobile Systems, Krakow, Poland (June 2009).
[23] Lane ND, Georgiev P, Qendro L. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In: ACM International Joint Conference on Pervasing and Ubiquitous Computing, Osaka, Japan.
[24] Apple ARKit, https://developer.apple.com/arkit/, [Accessed January 7, 2019].
[25] Ikea place, https://itunes.apple.com/app/ikea-place/id1279244498 [accessed: 2019-07-01].
[26] Sirvo llc, housecraft, https://itunes.apple.com/app/housecraft/id1261483849 [accessed: 2019-07-01].
[27] Sennheiser AMBEO Smart Headset, https://en-uk.sennheiser.com/finalstop [accessed January 7, 2019].
[28] Hong J, He J, Lam B, Gupta R, Gan W-S. Spatial audio for soundscape design: recording and reproduction. Appl Sci 2017;7(6):2017 (Jun 2017).
[29] Kınayoğlu G. Using audio-augmented reality to assess the role of soundscape in environmental perception. In: International Conference on Education and Research in Computer Aided Architectural Design in Europe, Instanbul, Turkey.
[30] Brown A, Kang J, Gjestland T. Towards standardization in soundscape preference assessment. Appl Acoust 2011;72(6):387–92. https://doi.org/10.1016/j.apacoust.2011.01.001.
[31] Merchan CI, Diaz-Balteiro L, Soliño M. Noise pollution in national parks: soundscape and economic valuation. Landscape Urban Plann 2014;123 (2014):1–9.
[32] Devos P. Soundecology indicators applied to urban soundscapes. In: Internoise; 2016 (August 2016).
[33] Aletta F, Kang J, Axelsson Ö. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. Landscape Urban Plann 2016;149(2016):65–74.
[34] Lundén P, Axelsson Ö, Hurtig M. On urban soundscape mapping: a computer can predict the outcome of soundscape assessments. Internoise August 2016;2016:4725–32.
[35] Cain R, Jennings P, Poxon J. The development and application of the emotional dimensions of a soundscape. Appl Acoust 2013;74(2):232–9. https://doi.org/10.1016/j.apacoust.2011.11.006 (Feb 2013).
[36] Barchiesi D, Giannoulis D, Stowell D, Plumbley MD. Acoustic scene classification: classifying environments from the sounds they produce. IEEE Signal Process Mag 2015.
[37] Core ML, https://developer.apple.com/documentation/coreml [accessed January 7, 2019].
[38] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(2011):2825–30.
[39] McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, Nieto O. Audio and music signal analysis in Python. In: Proc. of the 14th Python in Science Conference (SciPy 2015) (2015).
[40] Brossier P, Tintamar, Müller E, Philippsen N, Seaver T, Fritz H, Cyclopsian, Alexander S, Williams J, Cowgill J, Cruz A. aubio/aubio: 0.4.8 (Nov. 2018). https://doi.org/10.5281/zenodo.1494152.
[41] Creating an immersive ar experience with audio, https://developer.apple.com/documentation/arkit/creating_an_immersive_ar_experience_with_audio [accessed January 7, 2019].
[42] Murphy E, King EA. Environmental Noise Pollution: Noise Mapping, Public Health, and Policy. Amsterdam: Elsevier; 2014.
[43] Ekici I, Bougdah H. A review of research on environmental noise barriers. Build Acoust 2003;10(4):289–323.
[44] International Standards Organisation, Acoustics – attenuation of sound during propagation outdoors, ISO 9613-2; 1996.
[45] Kasten EP, Gage SH, Fox J, Joo W. The remote environmental assessment laboratory's acoustic library: an archive for studying soundscape ecology. Ecol Inf 2012;12(2012):50–67.
[46] D'Agostino RB, Belanger A, D'Agostino Jr Ralph B. A suggestion for using powerful and informative tests of normality. Am Stat 1990;44(4):316–21.
[47] mh Acoustics, em32 Eigenmike® microphone array release notes, mh acoustics, 25 Summit Ave, Summit, NJ 07901 (April 2013). prefix https://mhacoustics.com/sites/default/files/EigenmikeReleaseNotesV15.pdf.
[48] Creating a persistent ar experience, https://developer.apple.com/documentation/arkit/creating_a_persistent_ar_experience [accessed January 7, 2019].